一种面向智能车联网的缺失数据估计新方法 *

张德干 [a, b]，张　婷 [a, b†]，高瑾馨 [a, b]

(天津理工大学 计算机科学与工程学院 a. 计算机视觉与系统教育部重点实验室；b. 智能计算及软件新技术天津市重点实验室，天津 300384)

摘　要：智能车联网通过大量的地面传感器收集的数据来获得有关交通状况的信息，所收集的数据通常具有不规则的空间和时间分辨率，数据丢失是面对智能车联网中的一个常见问题。鉴于此，考虑了大型和多样化车联网中的缺失数据问题。通过在智能车联网中提取公共交通模式，比较了函数估计和张量分解等方法来估计这些缺失值的优劣后，提出了张量低秩近似估计新方法，该方法在缺失数据的情况下获得流量模式，得到大规模车联路网的低秩表示。通过不同的道路车联网实验测试，表明新方法的估计精度、数据集的偏差达到了较好的效果。

关键词：车联网；智能；数据丢失；估计；偏差

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.04.0319

New approach of losing data evaluation for IIOV

Zhang Degan[a, b], Zhang Ting[a, b†], Gao Jinxin[a, b]

(*a. Key Laboratory of Computer Vision & System of Ministry of Education, b. Tianjin Key Laboratory of Intelligent Computing & Novel Software Technology, School of Computer Science & Engineering, Tianjin University of Technology, Tianjin* 300384, *China*)

Abstract: Intelligent Internet of Vehicle (IIOV) gathers relative traffic information by all kinds of on-ground sensors. The gathered data often include irregular spatial and temporal resolution, so losing data is a common problem of IIOV. In order to solve this problem, this paper proposed a kind of new approach of losing data evaluation for IIOV which was named tensor low-rank approximation(VBPCA) based on the extracting the common traffic pattern and comparing the function estimation & tensor decomposition. The approach can get the traffic patterns under the cases of losing data and the expression of low-rank. In the experiments to test the approach, it select about 1000 road segments to do the analysis. The results show that this approach has good performance on evaluation accuracy, the bias of the data set, so it is very useful for the application of intelligent internet of vehicle.

Key words: internet of vehicle (IOV); intelligent; data losing; evaluation; bias

0　引言

随着传感器技术的进步，智能车联网现在可以从范围广泛的固定和移动传感器[1~3]收集交通数据。固定传感器如线圈检测器和路边的相机的空间范围往往有限，而移动传感器，如 GPS 探头收集的数据具有高度不稳定的空间和时间分辨率。这些问题造成了流量数据集中不可避免的数据丢失问题。此外，诸如检测器故障和有损通信系统等故障也可能导致交通信息不完整[4~8]。这可能会导致出现高比例的数据丢失。因此，丢失的数据是交通数据集中常见的问题[9~12]。在这方面的不同研究报告说，丢失的数据百分比可高达 90% [13]。对于交通管理系统，这是一个关键问题[14-16]。

解决数据缺失问题的方法大致可分为两类：函数估计和矩阵/张量完备化。在第一种情况下，通常假定缺失数据的问题局限于某些已知链接和时间间隔。这样，历史数据就可以用来获取目标道路与其邻近或过去道路之间的关系函数。例如，文献[17,18]利用历史数据建立相邻环路检测器之间的关系模型。这种关系函数是用来归结故障探测器缺失值，文献[19,20]]训练神经网络并利用时间特征估计缺失值。文献[21~23]还使用了类似的方法和应用最小二乘支持向量机估计缺失值。函数估计技术

收稿日期：2018-04-15；**修回日期：**2018-06-04　**基金项目：**国家自然科学基金资助项目（61571328）；天津市重大科技专项资助项目（15ZXDSGX00050，16ZXFWGX00010）；天津市科技支撑重点项目（17YFZCGX00360）；天津市自然科学基金资助项目（15JCYBJC46500）；天津市科技创新团队项目（12-5016，2015-23，13-5025）

作者简介：张德干（1970-），男，教授，博导，博士，主要研究方向为物联网、移动计算等；张婷（1972-），女（通信作者），河北唐山人，博士研究生，主要研究方向为车联网等（gandegande@126.com）；高瑾馨（1995-），女，硕士研究生，主要研究方向为网络通信等.

需要完整的历史数据来得到关系模型。因此，如果历史数据有缺少值，这些方法将无法使用。在实际场景中，未损坏的历史数据可能不可用。另一方面，矩阵和张量补全方法不需要训练数据来执行插补。因此，这些方法在交通研究领域获得了很大的收益。

相邻道路的交通状态趋向于强相关[24~26]。这些关系意味着道路网络可以用低维模型来表示。矩阵和张量完成方法利用这些模式来估计缺失值，通过获得不完全张量/矩阵的合适的低秩逼近。然而，以往关于交通数据集的矩阵/张量补全方法的研究大多集中在从几条道路或交叉口获得的数据上。例如，文献[24~26]用贝叶斯主成分分析（BPCA）对交通流数据进行插补。他们分析了一个由 100 条道路组成的小网络。文献[8~13]用张量分解方法进行缺失数据插补。通过分析，他们认为四个路段和其代表的数据从每个道路得到 3 张量。

城市规模网络中的交通状况也往往具有某些共同的全球模式。尽管方法有限，一些研究[21-26]已经考虑了大型网络中缺失数据的问题。这些研究没有分析不同道路类型（高速公路、干道、道路）和一周或半个月内不同日子的插补算法的性能。此外，他们没有分析估算插补交通数据的偏差和方差。总之，插补函数估计方法的应用受限于依赖未损坏的历史数据的大型网络。以前应用矩阵和张量完成方法的研究大多只考虑一个或几个交叉点的数据。这些研究通常不分析不同插补道路类型和一周内不同日子的性能。此外，还需要考虑方差、偏差以及低维模型的等级对插补性能的影响。

鉴于此，本文突破上述的限制，将执行缺失的数据插补的大型公路网扩展到高速公路、干线公路、次干线道路和支路等情形。提出张量低秩近似估计新方法（VBPCA），可以从不完全数据中提取全局流量模式。将上述方法与加权最小二乘法LS和近似奇异值分解 (FPCA)等方法的性能进行比较。分析这些方法针对不同道路类别和每周天数的性能、以及这些方法在估算速度数据中的方差和偏差。

本文的主要贡献如下：针对大规模智能车联网系统中缺失数据的问题，通过在智能车联网中提取公共交通模式，比较了函数估计和张量分解等方法来估计这些缺失值的优劣后，提出了张量低秩近似估计新方法（VBPCA），该方法在缺失数据的情况下获得流量模式，得到大规模路网的低秩表示。结果表明，该算法的性能对交通数据的日变化不敏感。此外，该方法还在加权相对误差（*WE*）、均方根误差（*MSE*）、方差（*V*）和偏差（*B*）等方面具有更好的性能。

1　车联网数据集与性能度量

针对车辆网数据集，本文用一组道路线段s_i的固定的E值来表示大小为p的测试道路网，例如$E=\{s_i\}_{i=1}^{p}$。在这项研究中，考虑平均速度数据。在区间链路s_i上的平均速度$(t_j-\Delta t, t_j)$由$z(s_i, t_j)$表示。采样间隔 Δt 是 3 min。每个链路s_i，创建了一个速度剖面$a_i\in R^{n}$，例如$a_i=[z(s_i,t_1),\ldots, z(s_i,t_n)]^T$。

速度配置文件包含每个链接的一天速度数据。根据这些速度曲线来获得网络配置矩阵$A\in R^{n\times p}$，比如$A=[a_1,\cdots,a_p]$。令$D\in R^{n\times p}$是相应的不完全观测数据矩阵。集Ω收录词条的位置在 *D* 的速度数据是可用的，集$\Theta=\Omega^C$表示在 *D* 丢失速度值的位置。

对张量的完成方法，创建的网络配置张量$\underline{A}\in R^{n\times p\times q}$，通过叠放在一起的网络配置矩阵$\{A_1, A_2,\cdots A_q\}$从不同的日子形成 3 张量。为此，使用 q＝7～14 天的数据。在这种情况下，不完全张量由$\underline{D}\in R^{n\times p\times q}$表示。

为了分析，考虑多个测试网络。每个网络中的道路属于天津的城市道路网，有足够的数据可供使用。每个测试网络的速度数据由天津的陆路交通管理局提供。

针对性能度量，下面描述不同的性能评估方法。

对于矩阵，定义加权相对误差（WE）实际 *A* 和估计$\hat{A}$之间的网络分布：

$$WE=\frac{\left\|W\otimes\left(A-\hat{A}\right)\right\|_F}{\left\|W\otimes A\right\|_F} \tag{1}$$

其中：符号 ⊗ 代表两个矩阵之间的元素相乘。矩阵$W\in R^{n\times p}$值的权重矩阵：

$$W_{ij}=\begin{cases}0 & (i,j)\in\Omega\\ 1 & (i,j)\in\Theta\end{cases} \tag{2}$$

矩阵$A\in R^{n\times p}$的 Fresenius 范数$\|A\|_F$定义为

$$\|A\|_F=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p}a_{ij}^2} \tag{3}$$

同样，定义为 *WE* 张量为

$$WE=\frac{\left\|\underline{W}\otimes\left(\underline{A}-\underline{\hat{A}}\right)\right\|_F}{\left\|\underline{W}\otimes\underline{A}\right\|_F} \tag{4}$$

其中：符号 ⊗ 代表两张量之间的元素相乘。张量$W\in R^{n\times p\times q}$为带权值的张量：

$$W_{ijk}=\begin{cases}0 & (i,j,k)\in\Omega\\ 1 & (i,j,k)\in\Theta\end{cases} \tag{5}$$

张量$\underline{A}\in R^{n\times p\times q}$的 Fresenius 范数$\|A\|_F$定义为

$$\|A\|_F=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}a_{ijk}^2} \tag{6}$$

加权相对误差通常被用来评估矩阵和张量完成算法的性能。计算均方根误差（MSE）的估计算法如下：

$$MSE_{mat}=\sqrt{\frac{1}{|\Theta|}\sum_{(i,j)\in\Theta}\left(a_{ij}-\hat{a}_{ij}\right)^2} \tag{7}$$

$$MSE_{ten}=\sqrt{\frac{1}{|\Theta|}\sum_{(i,j,k)\in\Theta}\left(a_{ijk}-\hat{a}_{ijk}\right)^2} \tag{8}$$

其中：$|\Theta|$代表集Θ的大小。计算估算速度数据中的偏差，如

下所示：

$$B_{mat}=\frac{1}{|\Theta|}\sum_{(i,j)\in\Theta}\left(a_{ij}-\hat{a}_{ij}\right) \quad (9)$$

$$B_{ten}=\frac{1}{|\Theta|}\sum_{(i,j,k)\in\Theta}\left(a_{ijk}-\hat{a}_{ijk}\right) \quad (10)$$

此外，计算了估计值的方差如下：

$$V_{mat}=\frac{1}{|\Theta|}\sum_{(i,j)\in\Theta}\left(a_{ij}-\bar{a}_{\Theta}\right)^2 \quad (11)$$

$$V_{ten}=\frac{1}{|\Theta|}\sum_{(i,j,k)\in\Theta}\left(a_{ij}-\bar{a}_{\Theta}\right)^2 \quad (12)$$

其中：$\bar{a}_{\Theta}$ 分别代表式（11）和（12）中 $\{\hat{a}_{ij}\}_{(i,j)\in\Theta}$ 和 $\{\hat{a}_{ijk}\}_{(i,j,k)\in\Theta}$ 的平均值。

2　缺失数据估计新方法

本章在讨论最小二乘法（LS）和固定点连续近似奇异值分解（FPCA）恢复不完整矩阵的丢失速度信息的补全算法的基础上，设计缺失数据估计的新方法（VBPCA）。

2.1　最小二乘法（LS）

在一个相互连接的网络中，速度等交通参数趋向相似，利用这些潜在的模式来恢复不完整矩阵 D 中的丢失的速度信息。为此，首先考虑完成网络配置的矩阵 A，运用主成分分析（PCA），可以从网络配置矩阵 A 中得到一个低秩近似（秩为 r）$\hat{A}=WX+M$，其中 $W\in R^{n\times r}$ 和 $X\in R^{r\times p}$ 是两个低秩矩阵，$M\in R^{n\times p}$ 是 A 的行的平均值，这个分解可通过求解下面的最小二乘优化问题得到：

$$\min_{\hat{A}}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(a_{ij}-\hat{a}_{ij}\right)^2,\quad \hat{a}_{ij}=W_i^T X_j+m_{ij} \quad (13)$$

在约束条件下，向量 $\{W_i\}_{i=1}^{r}$ 保持正交。就不完整矩阵 D 来说　，可以将问题用只观测速度数据 $\{d_{ij}\}_{(i,j)\in\Omega}$ 的重建误差最小化的形式来再次表示，其中 d_{ij} 代表在路段 s_j 时间 t_i 处速度值。因此，优化问题将成为

$$\min_{\hat{A}}\sum_{(i,j)\in\Omega}\left(d_{ij}-\hat{a}_{ij}\right)^2,\quad \hat{a}_{ij}=W_i^T X_j+m_{ij} \quad (14)$$

在本文中，用常用的梯度下降算法解决（14）中的最优化问题。

2.2　固定点连续近似奇异值分解

本节中讨论了另一种可选择的方法来估计丢失的交通信息。本文的目标是利用不同道路 $\{S_i\}_{i=1}^{p}$ 上的共同交通行为，在不完全数据矩阵 D 中恢复这些缺失的速度值。为此，需要得到一个合适的从不完整速度数据逼近的低秩矩阵 $\hat{A}$。此外，估计网络配置 $\hat{A}$ 也应该保护速度信息在具有一定的耐受极限 ε 的不完全数据矩阵 D 中可用，如 $\{|\hat{a}_{ij}-d_{ij}|<\varepsilon\}_{(i,j)\in\Omega}$。因此，可以设置如下的优化问题：

$$\min rank\left(\hat{A}\right)\ s.t.\left|\hat{a}_{ij}-d_{ij}\right|<\varepsilon,\forall(i,j)\in\Omega \quad (15)$$

上述提到的优化问题试图恢复带有最小数目的潜在成分的丢失速度数据同时保持所观察到的数据提供的速度信息 $\{d_{ij}\}_{(i,j)\in\Omega}$。然而，这是一个非凸和 NP 困难问题。为了使问题易于处理的，可以通过其凸包络替换 $rank(\hat{A})$，这原来是估计矩阵 $\hat{A}$ 的一个核范数 $\|\hat{A}\|_*$。这样，式（15）中的问题可以改写为

$$\min\|\hat{A}\|_*\ s.t.\left|\hat{a}_{ij}-d_{ij}\right|<\varepsilon,\forall(i,j)\in\Omega \quad (16)$$

其中：秩为 r 的矩阵 $\hat{A}$ 的核范数定义为 $\|\hat{A}\|_*=\sum_{i=1}^{r}\sigma_i$，$\sigma_i$ 是第 i 个奇异值矩阵。考虑固定点连续近似奇异值分解（FPCA）解决式（16）中定义的优化问题。

2.3　缺失数据估计的新方法

在上文讨论最小二乘法（LS）和固定点连续近似奇异值分解（FPCA）恢复不完整矩阵的丢失速度信息的补全算法的基础上，本节设计缺失数据估计的新方法，即张量低秩近似估计方法（VBPCA）。上文已经讨论了不同的矩阵补全方法来提取道路网络中的底层交通模式，然而，这些方法不能有效地利用交通数据集中的多路径依赖关系。例如，考虑一周中不同时间道路交通的行为。当然，交通参数，如速度，往往遵循类似的日常模式。可以通过创建交通数据的多路结构这种更有效的方式提取这些时间关系，为此，用 3 张量 $\underline{A}\in R^{n\times p\times q}$ 的形式代表速度数据。这个张量分布是由堆叠在一起的从不同时间中获得的网络分布矩阵 $\{A_1,A_2,\ldots,A_q\}$ 得到的。典范（CP）分解通常用于获得张量的低秩近似。对于不完整张量配置 $\underline{D}$，可以通过以下方式对观测速度数据进行重建误差最小化，得到一个合适的低秩近似 $\underline{\hat{A}}$：

$$\min_{\hat{A}}\frac{1}{2}\left\|\underline{W}\otimes\left(\underline{D}-\underline{\hat{A}}\right)\right\|_F^2,\quad \underline{\hat{A}}=\sum_{i=1}^{r}b_i^{(1)}\odot b_i^{(2)}\odot b_i^{(3)} \quad (17)$$

其中：$b_i^{(m)}$ 是现代因子矩阵 $B^{(m)}$ 的第 i 列向量。在式（17）中，符号 $\odot$ 表示矢量外积，而符号 $\otimes$ 代表两个张量之间元素相乘。因子矩阵 $B^{(1)}$、$B^{(2)}$ 和 $B^{(3)}$ 包含张量的不同模式下的公共通信模式。这些模式包括不同天和不同道路之间的公共交通行为。

应用 CP 加权优化（CP-OPT）从不完整的网络配置张量 $\underline{D}$ 中获得合适的估计 $\underline{\hat{A}}$。在展开张量方面利用 CP-OPT 来研究插补性能多方表征的影响。为此，通过组合多天的速度数据创建另一个网络概要矩阵 $U\in R^{n\times pq}|U=\left[A_1,\ldots,A_q\right]$。这种网络配置矩阵 U 本质上是网络配置张量 $\underline{A}$ 的展开表示，在这种情况下，相应的不完全数据矩阵 D_u 由表示。通过对观测速度数据的重建误差最小化，得到了不完全速度数据 D_u 的矩阵 U 的低秩逼近

$\hat{U}$：

$$\min_{\hat{U}} \frac{1}{2}\left\| W \otimes \left(D_u - \hat{U} \right) \right\|_F^2, \hat{U} = \sum_{i=1}^{r} b_i^{(1)} \odot b_i^{(2)} \qquad (18)$$

由此，运用 CP-OPT 来得到估计网络分布矩阵 U，这一方法称为张量低秩近似估计方法（VBPCA）。

基于上文的分析，缺失数据估计的新方法(VBPCA)的步骤可描述如下：

a)设置初始参数值。用一组道路线段s_i的固定的 E 值来表示大小为 p 的测试道路网，$E = \{s_i\}_{i=1}^{p}$。在区间链路s_i上的平均速度$(t_j - \Delta t, t_j)$由$Z(s_i, t_j)$表示。设置采样间隔 Δt 是 5 min。每个链路s_i，创建了一个速度剖面$a_i \in R^n$，

b)计算如下相关向量$a_i = [Z(s_i, t_1), \ldots, Z(s_i, t_n)]^T$。速度配置文件包含每个链接的一天速度数据。

c)通过车联网传感器网络采集不同日期不同道路类型（高速公路、主干道、快速路、次干道等）中的相关参数数据（如速度、时间、道路、车辆数等）得到数据集。从数据集中抽取出网络配置矩阵$A \in R^{n \times p}$，$A = [a_1, \ldots, a_p]$。令数据矩阵$D \in R^{n \times p}$是相应的不完全观测数据矩阵。

d)利用上文定义的式(1)~(6)来处理张量和范数，同时利用上文定义的式(7)~(8)来计算均方根误差，并且利用上文定义的式(9)~(12)来计算估计值的方差。

e)基于上文的式(17)对观测速度数据进行重建误差处理，得到一个合适的低秩近似$\hat{A}$， 应用 CP 加权优化（CP-OPT）从不完整的网络配置张量$\underline{D}$中获得合适的估计$\hat{\underline{A}}$。

f)利用上文的式(18)得到了不完全速度数据D_u的矩阵U的低秩逼近$\hat{U}$，从而得到估计网络分布矩阵 U。

g)基于测试数据集，评估考虑不同数量的潜在因素（秩）的缺失数据加权相对误差和不同道路网在一段时间内不同日子的算法加权相对误差。如果缺失数据估计精度的误差在预定的容忍范围内，则结束估计过程；否则，重返步骤 b)进行下一轮的估计，直到符合要求时为止。

上述算法的步骤可用伪代码将要描述如下：

1）set $E = \{s_i\}_{i=1}^{p}$，Δt =300s，$a_i = [Z(s_i, t_1), \ldots, Z(s_i, t_n)]^T$，$a_i \in R^n$

2）extract $A = [a_1, \ldots, a_p]$，$A \in R^{n \times p}$，$D \in R^{n \times p}$

3）calculate *WE*, ||*A*||, *MSE*, *B*, *V* by equations (1)~(12)

4）get $\hat{\underline{A}} = \sum_{i=1}^{r} b_i^{(1)} \odot b_i^{(2)} \odot b_i^{(3)}$ by equation (17)

5）get $\hat{U} = \sum_{i=1}^{r} b_i^{(1)} \odot b_i^{(2)}$ by equation (18)

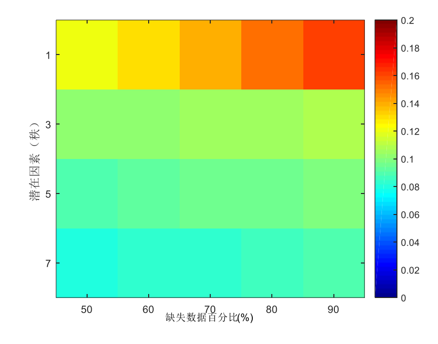
6）if error $E <$ threshold δ then exit or quit else go to 2).

3　测试与讨论

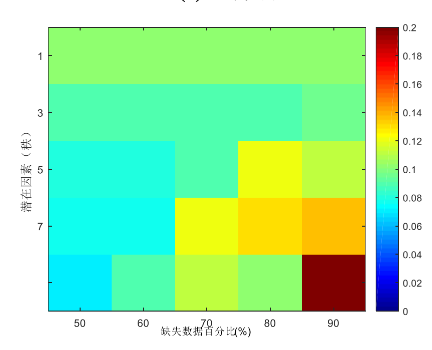
本章讨论秩（潜在因素的数量）的选择对算法性能的影响。图 1 显示了从京津高速公路获得的速度数据选择等级引起的不同算法的重建性能的变化。图 2 显示了从京津公路主干道获得的速度数据中的这些变化。讨论 LS、FPGA 和 VBPCA 这三种算法的性能，试图通过对观测到的速度信息的均方误差最小化来提取数据中的公共模式。

对于大量的缺失数据，这些算法的重建误差可以根据秩的选择而显著变化。此外，与高速公路相比，这些算法在干道上的波动性能更为显著（参见图 1 和 2）。图 1 和 2 中右侧图谱颜色和数值代表的意义是加权相对误差(含义与图 3 中纵坐标的含义相同)。图 1 和 2 中图谱颜色变化越频繁或颜色越深，表明性能波动越显著，相对误差越大。另一方面，对 FPGA 和 VBPCA 重构误差在不同的秩值处变化不显著。

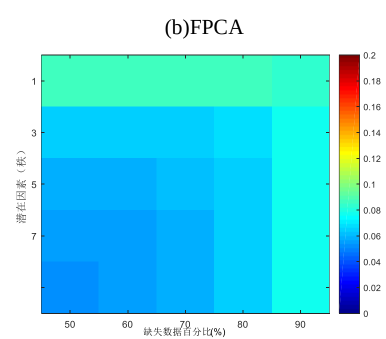
从图中可以看出，设计的 VBPCA 方法可以自动选择最优数量的因素且可以在不完全数据矩阵 D 中估计缺失值。图 1（c）和图 2（c）为 VBPCA 的秩值代表对可以用来重建估计网络分布矩阵的因素的最大数量的极限；同时，设定 VBPCA 对潜在因素最大限制的影响上限，鉴于此，可以得出这样的结论：如果等级合适的临界值不可用，VBPCA 也不会出现过拟合的现象。



(a)LS 方法



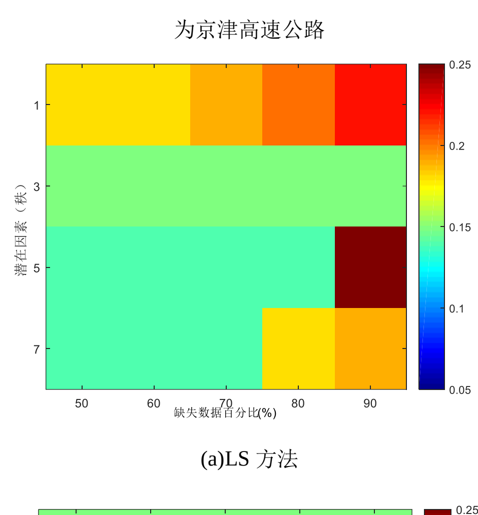
(b)FPCA



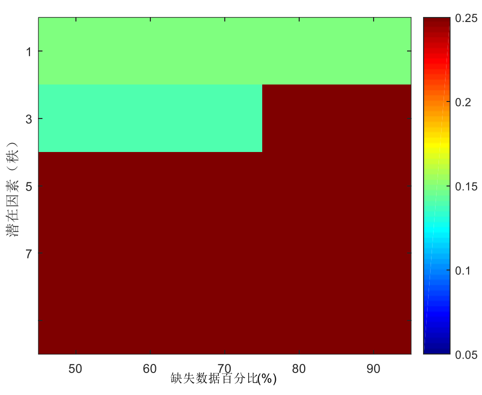
(c)VBPCA

图 1　考虑不同数量的潜在因素(秩)的缺失数据加权相对误差。测试网络

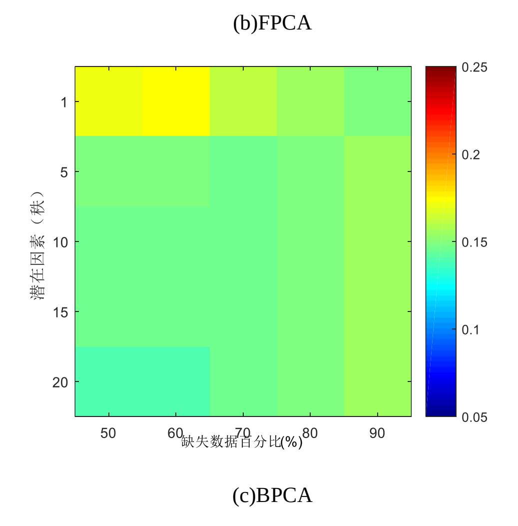
为京津高速公路



(a)LS 方法



(b)FPCA



(c)BPCA

图 2 考虑不同数量的潜在因素(秩)的缺失数据加权相对误差。测试网络为京津公路主干道

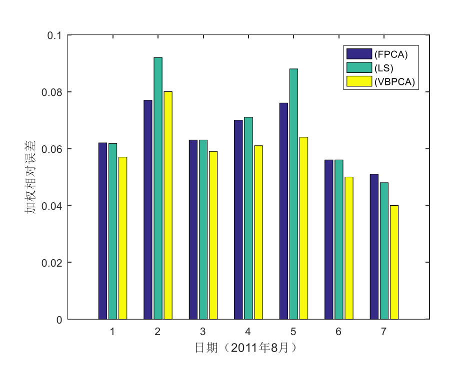
分析了各种道路网络下相关方法的性能情况,首先分析一下不同的插补方法在一周内的表现。图 3 显示了这些方法对不同道路类型的插补精度。在高速公路的情景下，VBPCA 达到最低的加权相对误差并紧随 FPCA 之后。与其他道路类别相比，高速公路的插补误差较之所有算法都要低。对于主要和次要干道，VBPCA 相比其他方法提供了更好的性能。对支路，VBPCA 也达到了更好的性能。在市区巷道的情况下，相关算法都存在较大的插补误差，但 VBPCA 的误差最小。

上述结果的原因可分析如下：LS、FPGA 和 VBPCA 这三种算法都试图通过寻找那些可以减少观测速度数据 $\{(d_{ij}-\hat{a}_{ij})^2\}_{(i,j)\in\Omega}$ 的重建误差平方和的交通模式来填补缺失值。基于最小二乘的方法中，多路径表示（张量法）趋向于达到最佳性能。此外，主干道和交流道、多路表示（张量方法）也达到了比其他方法如 FPGA 和 VBPCA 更好的插补精度。然而在高速公路方面，考虑多路表示的优势并不明显。看来，张量表示在小地方道路交通行为更不稳定的情况下更有用。在这种情况下，速度数据的多路表示是提取底层流量模式的一种有效方法。

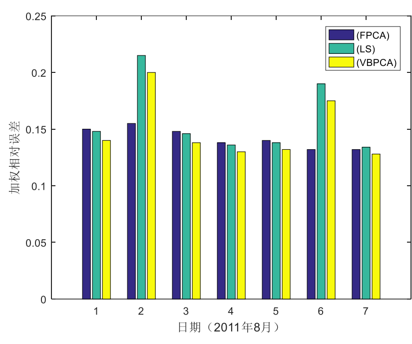
图 3 所示为 LS、FPGA 和 VBPCA 这三种算法在一周的不同时间的插补误差。结果显示一周时间内不同道路类别的速度数据。在这种情况下，丢失的数据百分比为 70%。对于高速公路，在大多数时间里 VBPCA 与其他方法相比具有较低的加权相对误差。正如预期的那样，VBPCA 对于高速公路的速度数据具有最低的总体估算误差。对于主干道，这三种方法在大部分时间内都有类似的性能。然而，在某些时间里，FPGA 和 LS 会具有较大的估计误差，但是 VBPCA 从一天到另一天的估计性能变化不显著。也在 LS、FPCA 和 VBPCA 对于快速路的性能上观察到类似的趋势。对于主要和局部次干道，所有这三种方法在七天内都产生了较大的插补错误。图 3 中不同日期下 VBPCA 算法效果有时并非最佳，造成这种现象的原因是重建数据出现了噪声或冗余，导致在去除噪声或冗余数据过程中插补性能受到了一定程度的扰动影响，这种现象在误差容忍范围内，因此是正常的。可以得出这样的结论：VBPCA 的插补性能与其他方法如 LS、FPCA 相比较，在交通条件每日变化的条件下具有更强的鲁棒性。

表 1 显示了所提出的各种方法在恢复速度数据中引起的偏差，偏置的单位是 km/hr。结果表明，在所有的测试用例中，当偏置值小于 1 km/hr 时，所提出的算法不会对估计数据添加显著的偏差。然而，由 FPCA 和 LS 得到的估计速度数据与其他方法得到的数据相比较有稍高的偏置（≈0.5 km/hr）。依此类推，可计算出不同道路类型的估计速度数据的方差，方差的单位是 km^2/hr^2。括号中的值代表估算速度数据相对于实际速度数据的百分比变化。它也显示了实际速度数据的方差，正如预期的那样，插补算法低估了估算数据的方差。例如，高速公路的速度数据的实际方差约为 149 km^2/hr^2。然而，不同的插补方法获得的速度数据的方差为 90–120 km^2/hr^2。此外，随着丢失数据百分比的增加，实际和预期的速度数据间的差异变得更大了。对于高速公路，VBPCA 提供了速度数据方差的最佳估计。对于其他类型的道路如主干道、快速路、次干道和巷道，由 VBPCA 得到的估算数据方差最接近实际速度数据的方差。

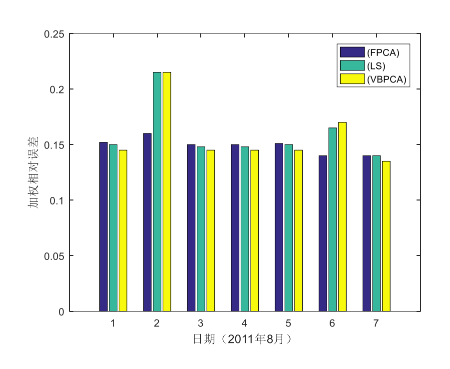
结果表明，VBPCA 方法的性能与 LS、FPCA 相比在秩的选择上高度敏感。VBPCA 由于它的性能对于日变化是最不敏感的，所以对于交通数据集的插补是特别有用的。此外，它为不同道路类别的其他算法提供更好的性能。



(a)京津高速公路



(b)京津公路主干道



(c)天津市区次干道

图 3 不同道路网在一周内不同日子的算法加权相对误差。重建误差是在 70%的数据丢失的情况下进行的

表 1 不同道路类型的估算速度偏差

道路类型	丢失数据	偏差		
		FPCA	LS	VBPCA
京津高速公路 平均速度=120 km/hr	10%	0.009	0.069	0.004
	20%	0.013	0.065	-0.007
	40%	0.009	0.068	-0.012
	60%	0.008	0.076	0.008
	90%	0.006	0.065	-0.002
公路主干道 平均速度=100 km/hr	10%	0.009	0.438	-0.003
	20%	0.009	0.409	0.006
	40%	0.005	0.468	0.016
	60%	-0.005	0.379	0.004
	90%	-0.007	0.378	0.005
市区快速路 平均速度=80 km/hr	10%	-0.009	0.479	0.0019
	20%	0.003	0.349	-0.007
	40%	0.002	0.418	-0.006
	60%	-0.005	0.419	0.008
	90%	-0.027	0.432	-0.007
市区次干道 平均速度=60 km/hr	10%	0.011	0.370	-0.013
	20%	0.011	0.385	-0.013
	40%	0.003	0.395	0.022
	60%	-0.003	0.418	0.013
	90%	-0.002	0.419	0.015

4 结束语

针对丢失数据这个车联网中面临的常见问题，本文比较了三种方法去估计在大型车联网的数据集中丢失的情形。为此，通过在智能车联网中提取公共交通模式，比较了函数估计和张量分解等方法来估计这些缺失值的优劣后，提出了张量低秩近似估计新方法，该方法在缺失数据的情况下获得流量模式，得到大规模路网的低秩表示。分析了不同类型的道路以及在一周或半个月内不同日子的各种矩阵和张量完成方法的重建精度；还分析了潜在因素的选择对恢复速度数据估计精度的影响、估计速度数据中的方差和偏差。通过不同的道路车联网实验测试，表明提出新方法的估计精度、数据集的偏差达到了较好的效果。

参考文献：

[1] Zhang Degan, Li Guang, Zheng Ke. An energy-balanced routing method based on forward-aware factor for wireless sensor network [J]. IEEE Trans on Industrial Informatics, 2014, 10 (1): 766-773.

[2] Liang Yanping. A kind of novel method of service-aware computing for uncertain mobile applications [J]. Mathematical and Computer Modelling, 2013, 57 (3-4): 344-356.

[3] Chen Jieqiong, Mao Guoqiang. Capacity of cooperative vehicular networks with infrastructure support: multi-user case [J]. IEEE Trans on Vehicular Technology, 2018, 67 (2): 1546-1560.

[4] Zheng Ke, Zhang Ting. A Novel Multicast Routing Method with Minimum

Transmission for WSN of Cloud Computing Service [J]. Soft Computing, 2015, 19 (7): 1817-1827.

[5] Song Xiaodong, Wang Xiang. New agent-based proactive migration method and system for big data environment (BDE) [J]. Engineering Computations, 2015, 32 (8): 2443-2466.

[6] Zhang Xiaodan. Design and implementation of embedded un-interruptible power supply system (EUPSS) for Web-based mobile application [J]. Enterprise Information Systems, 2012, 6 (4): 473-489.

[7] Zhang Degan. A new approach and system for attentive mobile learning based on seamless migration [J]. Applied Intelligence, 2012, 36 (1): 75-89.

[8] Song Xiaodong, Wang Xiang. Extended AODV routing method based on distributed minimum transmission (DMT) for WSN [J]. International Journal of Electronics and Communications, 2015, 69 (1): 371-381.

[9] Zheng Ke. Novel quick start (QS) method for optimization of TCP [J]. Wireless Networks, 2016, 22 (1): 211-222.

[10] Ma Zhen. New AODV routing method for mobile wireless mesh network (MWMN) [J]. Intelligent Automation and Soft Computing, 2016, 22 (3): 431-438.

[11] Zhou Shan, Tang Yameng. A low duty cycle efficient MAC protocol based on self-adaption and predictive strategy [J]. Mobile Networks and Applications, 2017, 2. (DOI: 10. 1007//s11036-017-0878-x)

[12] Liu Si, Zhang Ting. Novel unequal clustering routing protocol considering energy balancing based on network partition & distance for mobile education [J]. Journal of Network and Computer Applications, 2017, 88 (15): 1-9.

[13] Wang Xiang, Song Xiaodong. New clustering routing method based on PECE for WSN [J]. EURASIP Journal on Wireless Communications and Networking, 2015, 2015 (162): 1-13.

[14] Wang Xiang, Song Xiaodong. New medical image fusion approach with coding based on SCD in wireless sensor network [J]. Journal of Electrical Engineering & Technology, 2015, 10 (6): 2384-2392.

[15] Wang Xiang, Song Xiaodong. A kind of novel VPF-based energy-balanced routing strategy for wireless mesh network [J]. International Journal of Communication Systems, 2017, 30 (6): 1-15.

[16] Zhu Yanan. A new constructing approach for a weighted topology of wireless sensor networks based on local-world theory for the Internet of things (IOT) [J]. Computers & Mathematics with Applications, 2012, 64 (5): 1044–1055.

[17] Zhang Degan, Wang Xiang, Song Xiaodong. A novel approach to mapped correlation of ID for RFID anti-collision [J]. IEEE Trans on Services Computing, 2014, 7 (4): 741-748.

[18] Ma Zhen. Shadow Detection of moving objects based on multisource information in Internet of things [J]. Journal of Experimental & Theoretical Artificial Intelligence, 2017, 29 (3): 649-661.

[19] Niu Hongli, Liu Si. Novel positioning service computing method for WSN [J]. Wireless Personal Communications, 2017, 92 (4): 1747–1769.

[20] Song Xiaodong, Wang Xiang. New agent-based proactive migration method and system for big data environment (BDE) [J]. Engineering Computations, 2015, 32 (8): 2443-2466.

[21] Li Wenbin, Liu Si. Novel fusion computing method for bio-medical image of WSN based on spherical coordinate [J]. Journal of Vibroengineering, 2016, 18 (1): 522-538.

[22] Zhou Shan, Chen Jie. New DV-distance method based on path for wireless sensor network [J]. Intelligent Automation and Soft Computing, 2017, 23 (2): 219-225.

[23] Ma Zhen. A novel compressive sensing method based on SVD sparse random measurement matrix in wireless sensor network [J]. Engineering Computations, 2016, 33 (8): 2448-2462.

[24] Zhao Chenpeng. A new medium access control protocol based on perceived data reliability and spatial correlation in wireless sensor network [J]. Computers & Electrical Engineering, 2012, 38 (3): 694-702.

[25] Zhou Shan, Chen Jie. New mixed adaptive detection algorithm for moving target with big data [J]. Journal of Vibroengineering, 2016, 18 (7): 4705-4719.

[26] Niu Hongli, Liu Si. Novel PEECR-based clustering routing approach [J]. Soft Computing, 2017, 21 (24): 7313-7323.